Math is Music – Stats is Literature

Or why are there no six year old novelists?

Dick De Veaux, Williams College

With thanks to:

Paul Velleman, Cornell University

Norean Sharpe, Babson College

Prodigies

- **Math, music, chess**
  - Gauss
    - Story of age 3 adding up 1 + … + 100
    - Magnum Opus *Disquisitiones Arithmeticae* by 21
  - Pascal
  - Mozart, Schubert, Mendelssohn
  - Bobby Fischer

- **Why these three areas?**
- **Each creates its own world with its own set of rules**
  - There is no "experience" required
  - Once you know the rules, you are free to create anything

Prodigies in Literature?

- **Thomas Dulack**
  - "There are no child prodigies in literature."
- **List from Wikipedia**
  - William Cullen Bryant
  - Thomas Chatterton
  - H.P. Lovecraft
  - Mattie Stepanek (Died at 13)
  - Lope de Vega
  - Henriett Seth-F.

- **Others?**
  - Mary Wollstonecraft Shelley
- **Why?**
  - Literature is about the world, not about rules. It deals with life's experience and the wisdom we develop over time.

Statistics – What do students find so hard?

- "Understood the material in class, but found it hard to do the homework"
- "Should be more like a math course, with everything laid out beforehand"
- "More problems in class should be like the HW and tests"
- "Say what we need to know and don't add anything else"

What is "easy"?

- **The math part – well, not "easy", but…**
  - Math is axiomatic – logical – laid out beforehand
  - Given one example, we can change the numbers and it still makes sense

PROBLEM 5 : A sheet of cardboard 3 ft. by 4 ft. will be made into a box by cutting equal-sized squares from each corner and folding up the four edges. What will be the dimensions of the box with largest volume ?

A Typical Statistics Problem

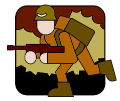
29. Insulin and diet. A study published in the *Journal of the American Medical Association* examined people to see if they showed any signs of IRS (insulin resistance syndrome) involving major risk factors for Type 2 diabetes and heart disease. Among 102 subjects who consumed dairy products more than 35 times per week, 24 were identified with IRS. In comparison, IRS was identified in 85 of 190 individuals with the lowest dairy consumption, fewer than 10 times per week.

a) Is this strong evidence that IRS risk is different in people who frequently consume dairy products than in those who do not?

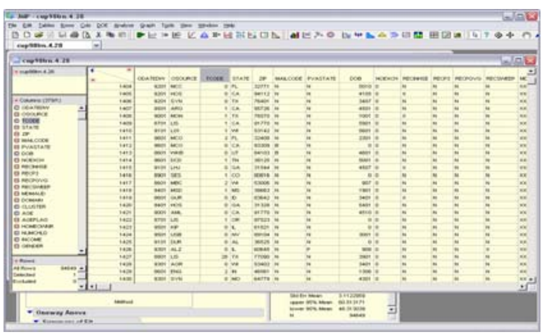
b) Does this prove that dairy consumption influences the development of IRS? Explain.

Paralyzed Veterans of America

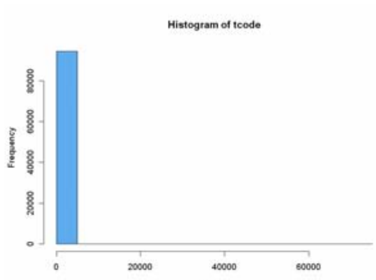
- **KDD 1998 cup**
- **Mailing list of 3.5 million potential donors**
- **100,000 customers as “training set”**
  - Predictors -- 481 variables
  - Responses -- In a recent campaign
    - Did they give?
    - If so, how much?
- **Based on this, is there a better strategy for who should get the next mailing?**



What’s “Hard”? -- Example

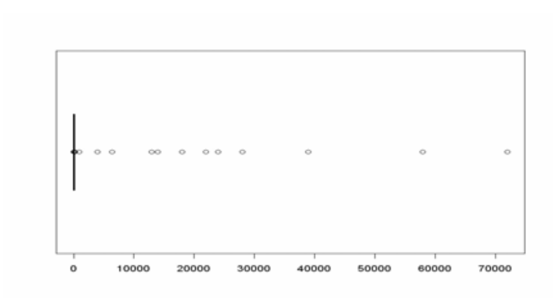


T-Code



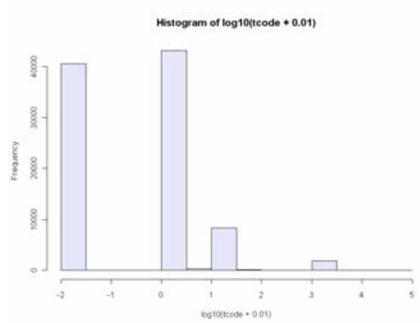
Max	**72000**
Q3	**2**
Median	**1**
Q1	**0**
Min	**0**
Mean	**54.41**
SD	**957.5**

More Tcode



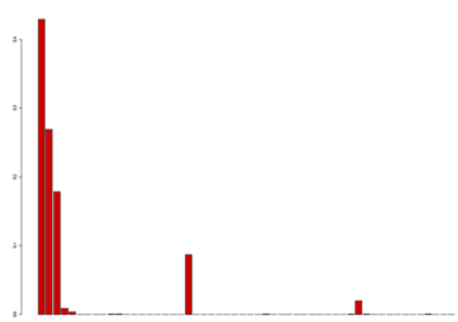
MAA Short Course

Transformation?



MAA Short Course

Categories?

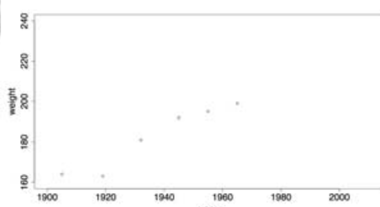


MAA Short Course

What does it mean?

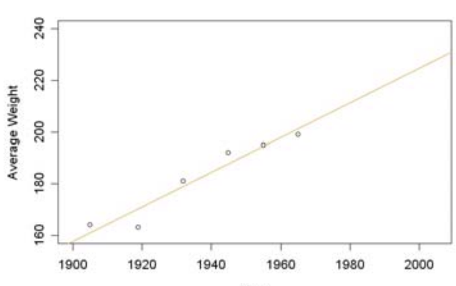
T-Code	Title						
0	_	16	DEAN	48	CORPORAL	109	LIC.
1	MR.	17	JUDGE	50	ELDER	111	SA.
1001	MESSRS.	17002	JUDGE & MRS.	56	MAYOR	114	DA.
1002	MR. & MRS.	18	MAJOR	59002	LIEUTENANT & MRS.	116	SR.
2	MRS.	18002	MAJOR & MRS.	62	LORD	117	SRA.
2002	MESDAMES	19	SENATOR	63	CARDINAL	118	SRTA.
3	MISS	20	GOVERNOR	64	FRIEND	120	YOUR MAJESTY
3003	MISSES	21002	SERGEANT & MRS.	65	FRIENDS	122	HIS HIGHNESS
4	DR.	22002	COLNEL & MRS.	68	ARCHDEACON	123	HER HIGHNESS
4002	DR. & MRS.	24	LIEUTENANT	69	CANON	124	COUNT
4004	DOCTORS	26	MONSIGNOR	70	BISHOP	125	LADY
5	MADAME	27	REVEREND	72002	REVEREND & MRS.	126	PRINCE
6	SERGEANT	28	MS.	73	PASTOR	127	PRINCESS
9	RABBI	28028	MSS.	75	ARCHBISHOP	128	CHIEF
10	PROFESSOR	29	BISHOP	85	SPECIALIST	129	BARON
10002	PROFESSOR & MRS.	31	AMBASSADOR	87	PRIVATE	130	SHEIK
10010	PROFESSORS	31002	AMBASSADOR & MRS	89	SEAMAN	131	PRINCE AND PRINCESS
11	ADMIRAL	33	CANTOR	90	AIRMAN	132	YOUR IMPERIAL MAJEST
11002	ADMIRAL & MRS.	36	BROTHER	91	JUSTICE	135	M. ET MME.
12	GENERAL	37	SIR	92	MR. JUSTICE	210	PROF.
12002	GENERAL & MRS.	38	COMMODORE	100	M.		
13	COLONEL	40	FATHER	103	MLLE.		
13002	COLONEL & MRS.	42	SISTER	104	CHANCELLOR		
14	CAPTAIN	43	PRESIDENT	106	REPRESENTATIVE		
14002	CAPTAIN & MRS.	44	MASTER	107	SECRETARY		
15	COMMANDER	46	MOTHER	108	LT. GOVERNOR		
15002	COMMANDER & MRS.	47	CHAPLAIN				

Sensible Model?



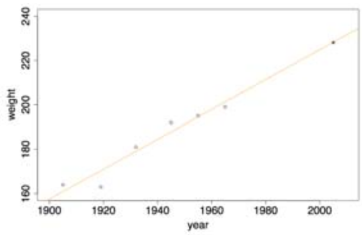
Linear Regression

$$\text{Predicted Weight} = -1121 + 0.6733 * \text{Year}$$



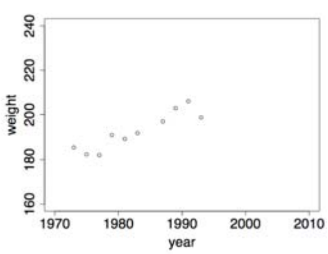
Forecast

$$Predicted\ Weight = -1121 + 0.6733 * 2005 = 228.21$$
$$Actual\ Weight\ 2005\ Team = 228.1\ lbs$$



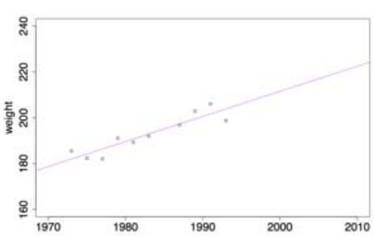
Williams College

- Div III “Powerhouse” -- Sears Cup 10/11 years

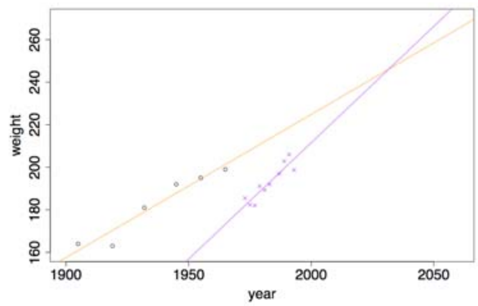


Williams Forecast

$$Predicted\ Weight = -1977.22 + 1.094 * Year$$
$$Predicted\ Weight\ 2005 = 216.97\ lbs.$$



Ephs Crush Longhorns?



What's the Hard part?

- **Putting everything together**
  - Real World
  - Does is make "sense"?
  - Which method to use?
- **When did Stats become hard?**
  - Roxy Peck's troublemakers

Teaching Calculus

- **Of course, it's not easy**
- **But (1st semester) Calculus has fewer concepts to get across**
  - Functions and Graphs (review)
  - Limits (review?)
  - Continuity (some review)
  - Derivative – max and min
  - Implicit differentiation
  - Antiderivative – area
  - Fundamental Theorem

Emphasis is on Computation

What about Statistics?

- **Exploratory Data Analysis**
  - Summarizing distributions/relationships
- **Data Collection**
  - Developing and implementing surveys
  - Interpreting surveys (errors)
  - Experimental Design – Causation vs. Correlation
  - Population vs. Sample
- **Randomness and Variation**
  - Random variables
  - Center and spread
- **Inference, confidence, and significance**
  - Confidence Interval
  - Hypothesis Testing – The Scientific Method !
- **Models and limits to models**
  - Residual analysis
  - Assumptions
  - Can we use the model?
- **Probability????**

Emphasis is on Interpretation

What's Hard? Seven Unnatural Acts

1.Think Critically

- **Know what we want to know.**
  - What's the QUESTION?
- **Challenge the data's credentials.**
- **Challenge how they were collected**
  - Look for bias
  - Have they ever done this in Calculus?
  - Is the cone "really" a cone?
- **Plot the data**
  - And ask about lurking variables

2. Be Skeptical

- **Being skeptical is part of critical thinking**
  - Be cautious about making claims based on data.
- **"Trust every analysis, but plot the residuals."**
  - Skeptical statisticians expect the unexpected, so we go looking for it.
- **Assumptions – limits to analysis**
- **Question the analysis** –
  - Not just is the answer correct, but
  - Is it appropriate?
  - Did it answer the question
  - Statistical vs. Practical significance

3. Think about Variation

- **Everyone find it easier to think about values rather than variation, but this is the main subject of our course**

Statistics is about Variation

Example

- **A town has two hospitals**
  - Large hospital about 100 babies a day
  - Smaller hospitals about 15 babies a day

- **Over the course of the year, which hospital (if either) would probably have more days in which more than 60% of the babies born are male?**

4. Focus on What We Don't Know

- **In most science and math courses, we focus on what we know**
- **Isn't that hard enough?**
- **Statisticians are a little strange**

Confidence Intervals

- **We don't say "The mean is 31.2".**
- **We don't say "The mean is probably 31.2"**
- **We don't say "The mean is close to 31.2".**
- **All we can manage is**
  - "*The mean is close to 31.2…. Probably*
  - *And we go on at great length to tell you how wrong we probably are*

5. Probability and Rare Events

- **Conditional, joint, rare events; randomness**
  - This is just plain hard.
- **It is easy to show that we don't naturally think clearly about conditional probabilities.**
  - But we need to in order to make rational decisions in the world

Linda (Tversky & Kahneman)

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and she participated in antinuclear demonstrations.

Rank these in order of Likelihood

Linda:

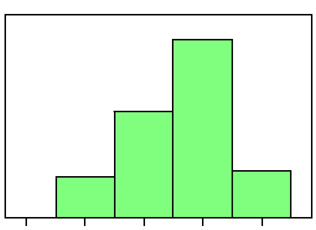
a) Is a teacher in an elementary school
b) Works in a bookstore and takes yoga classes.
c) Is active in the feminist movement.
d) Is a psychiatric social worker
e) Is a member of the League of Women Voters.
f) Is a bank teller.
g) Is an insurance salesperson.
h) Is a bank teller who is active in the feminist movement.

$$\Pr(A) \geq \Pr(A \wedge B) \leq \Pr(B)$$

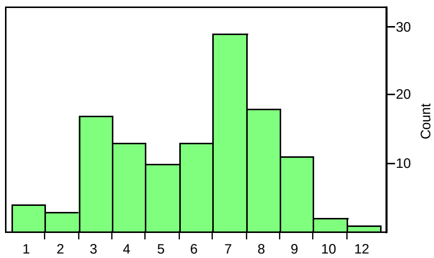
Pick a number at Random

1 2 3 4

Random?



Random II



6. Solution as Process

- **In Statistics, there is often not a simple right answer, but a process of inquiry**
- **Induction → Deduction → Induction**
- **The language of science**

7. Embrace Vague Concepts

- **Deal upfront with imprecise concepts**
  - Skewed vs. symmetric
  - Center, spread
  - Unimodal or not?
  - Are the assumptions and conditions met?
    - What impact does it have on the *decision?*

Statistics as Problem Solving

- **Unless you know what the problem is, don't start the analysis**
  - Don't even collect the data
    - CEO at First USA
    - Xerox
- **Identify the data you have**
  - Know the source
  - Identify the W's of the data

Clearing the woods

- **Trim out unnecessary topics**
  - How to choose bin widths
  - Formulas for grouped data
  - Shortcut formulas
  - Testing mean, sigma known
  - Combinatorics
  - Probability?

Models - Honesty is the best policy

- **Be honest about models**
  - Tell them Statistics is really about models
  - A model is a simplification of reality.
- **We know the model's not perfect**
  - So be sure to check if it's appropriate!

All Models are Wrong…

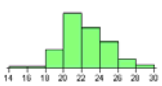
George Box:

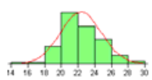
"All models are wrong… but some are useful"

"Statisticians, like artists, have the bad habit of falling in love with their models"

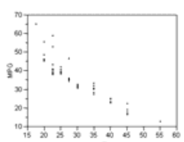
Common Models

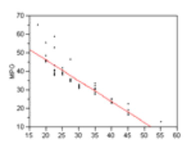
- Probability models





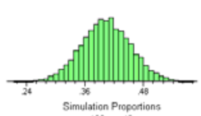
- Regression model





Common Models

- **Simulation – by "hand" or computer**



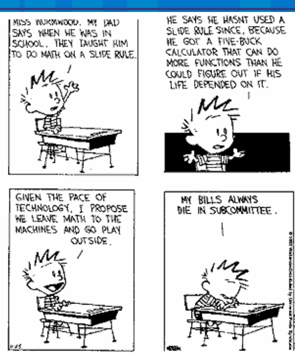
“Pay Dirt” Models

- **Sampling distribution models**
  - By now students know that models are idealized
  - They’ve seen probability models and simulations: CLT follows naturally
- **Null hypothesis models**
  - Wrong (we hope) but useful

Technology frees us

- **Calculation is for calculators and statistics packages.**
- **Let the technology do the work, so students can think about statistical thinking.**
- **Let them do it so we can “play Statistics”**

Play Stats



More Help – Reality Checks

- **Emphasize the concepts over the formulas.**
  - The answer is wrong if it makes no sense -- even if you pushed the buttons you meant to push or gave the command you intended
- **Check that the results are plausible**

9. Professors. A friend tells you about a recent study dealing with the number of years of teaching experience among current college professors. He remembers the mean but can't recall whether the standard deviation was 6 months, 6 years, or 16 years. Tell him which one it must have been, and why.

Making Statistics Relevant

- **Emphasize that Statistics is problem solving for science and industry**
  - Every business decision involves statistics
  - Most junk mail is marketing "experiments"

Emphasizing Communication

- **Machines are better at computing than even my best math students**
- **We take a sample of commuting times and find the mean to be 31.7 with a 95% confidence interval of (24.6, 38.8) minutes.**
  - What does this mean?
  - What doesn't this mean?

What it doesn't mean

95% *of all students* who commute have commute times between 24.6 and 38.8 minutes

We are 95% confident that *a randomly selected student who commutes* will have a commute times between 24.6 and 38.8 minutes

Mean commute time is 31.7 minutes 95% *of the time*.

95% *of all samples* will have mean commute times between 24.6 and 38.8 minutes

How can we help

- **Give them an outline for putting the real world into a framework (Deming)**
  - What's the problem? (Plan)
  - What are the mechanics? (Do)
  - What have we learned? (Report)
  - What next? (Act)

GAISE: Guidelines for Assessment and Instruction in Statistics Education[1]

- **Emphasize statistical literacy and develop statistical thinking**
- **Use real data**
- **Stress conceptual understanding rather than mere knowledge of procedures**
- **Foster active learning**
- **Use technology to develop conceptual understanding and analyze data**
- **Use assessments to improve and evaluate learning**

[1]GAISE Report (2005), Members of the GAISE Group: Martha Aliaga, George Cobb, Carolyn Cuff, Joan Garfield (Chair), Rob Gould, Robin Lock, Tom Moore, Allan Rossman, Bob Stephenson, Jessica Utts, Paul Velleman, and Jeff Witmer

Take Home Messages

- **Stats is about the real world**
  - It's messy – literature not music
  - Math is important – but it's not the message
  - Motivate by rooting the course in examples and real data that's relevant to students
  - Tell the story of Statistics so students take home a complete picture, not a set of tools
  - Technology frees the student to *think* about the world
  - We need to give the student a structure for a chaotic world (Deming)
    - Make them better problem solvers
  - Help them with unnatural thinking

Intro Stats – For liberal arts

Get them involved

- **Talk about them**

How many siblings do you have?
How would you describe yourself on the following political scale?

Sex (M/F)
Your class (Frosh/Soph/Jun/Sen)
Do you believe in God?
Pick a random number between 1 and 10:
How tall are you (in inches)?
How much do you weigh (in pounds)?
How many people have you dated in the past 6 months?
How many Facebook friends do you have?
How many alcoholic drinks did you have last night?
Do you play a varsity sport? (Or will try out?)
What's your favorite band?
Who do you support in the November Presidential election?
How many people do you know with the following last names?

How many songs are on your iPod?
What kind of food do you find most appealing?
What kind of music do you most enjoy?

Where are we going?

- **Data Exploration**
  - Data types – why?
  - Data displays and summaries
  - Data collection (later)
  - What questions?
- **Relationships between categorical variables**
- **Groups**
  - Relationship between quantitative and categorical

What next?

- **Relationships between quantitative variables**
  - How to summarize?
  - History of correlation and regression
- **Data collection**
  - Surveys
  - Observational studies
  - Experiments
- **What would we like to know?**

Probability

- **Historically probability was first**
  - Equally likely events
  - What do we mean?
    - Law of large numbers
    - Rules
  - Addition rule
  - Multiplication rule
  - Independence
  - Conditioning
  - Bayes?

Inference

- **The black swan**
  - Does seeing 1,000,000 white swans out of 1,000,000 prove that “all swans are white”?
  - How do we prove that?
- **Statistical inference is the uncertain version of the black swan**
  - I think I saw a black swan. If all swans were white, this would be really unlikely. Therefore….

Simulation

- **Suppose we had a fair coin**
  - Pennies from the 1960’s
  - When do we decide the coin is biased?
- **How do we know how the proportion is supposed to behave?**
  - Simulation
  - Resampling
  - Probability
  - Central Limit Theorem

Inference

- **Hypothesis test**
  - Two types of errors
  - Decisions under uncertainty
  - What is a P-value?
  - What do I do if the P-value is small? Large?
- **Confidence interval**
  - Why bother?
  - What does it mean?
- **Power & Effect size**

Intro Stats – Stat 201

What's Different?

- **Students have had calculus**
- **More important, they are comfortable with:**
  - Formulas
  - Abstraction
  - Computers
- **Goals -- Same as 101, plus:**
  - Programming
  - Probability
  - Derivations (at least seeing them)

Successful Data Mining in Practice

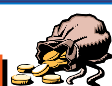
Richard D. De Veaux

Williams College

January 4, 2009

deveaux@williams.edu

Reason for Data Mining



Data = $$

Data Mining Is…

"the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." --- Fayyad

"finding interesting structure (patterns, statistical models, relationships) in data bases".--- Fayyad, Chaduri and Bradley

"a knowledge discovery process of extracting previously unknown, actionable information from very large data bases"--- Zornes

" a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions." ---Edelstein

Data Mining Models – a partial list

- **Traditional statistical models**
  - Linear regression, logistic regression, splines, smoothers etc.
  - Vendors are adding these to DM software
- **Visualization Methods**
- **Neural networks**
- **Decision trees**
- **K Nearest Neighbor Methods**
- **K-means**

What makes Data Mining Different?

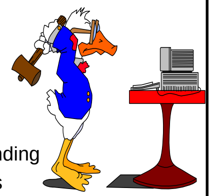
- **Massive amounts of data**
  - Number of rows (cases)
  - Number of columns (variables)
- **UPS**
  - 16TB – U.S. library of congress
  - Mostly tracking
- **Google**
  - 1 PB every 72 minutes
- **Low signal to noise**
  - Many irrelevant variables
  - Subtle relationships
  - Variation

Why Is Data Mining Taking Off Now?

- **Because we can**
  - Computer power
  - The price of digital storage is near zero
- **Data warehouses already built**
  - Companies want return on data investment

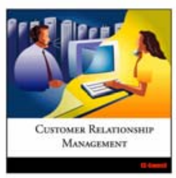
Users are Also Different

- **Users**
  - Domain experts, not statisticians
  - Have too much data
  - Want *automatic* methods
  - Want useful information without spending all their time doing statistical analysis



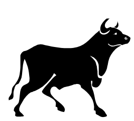
Customer Relationship Management

- **Transactional Data**
  - Customer retention
  - Upselling opportunities
  - Customer optimization across different areas
- **Marketing Experiments**
  - Often, new hypotheses are generated by data mining a planned experiment.
  - Segmentation



Financial Applications

- **Credit assessment**
  - Is this loan application a good credit risk?
  - Who is likely to declare bankruptcy?
- **Financial performance**
  - What should be a portfolio product mix



Manufacturing Applications

- **Product reliability and quality control**
- **Process control**
  - What can I do to improve batch yields?
- **Warranty analysis**
  - Product problems
  - Service assessment
  - Adverse experiences – link to production

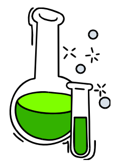
Medical Applications

- **Medical procedure effectiveness**
  - Who are good candidates for surgery?
- **Physician effectiveness**
  - Which tests are ineffective?
- **Which physicians are likely to over-prescribe treatments?**
  - What combinations of tests are most effective?

E-commerce

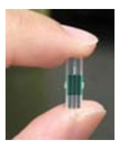
- **Automatic web page design**
- **Recommendations for new purchases**
- **Cross selling**

Pharmaceutical Applications



- **Combine clinical trial results with extensive medical/demographic information**
- **Non traditional uses of clinical trial data warehouse to explore:**
  - Prediction of adverse experiences – combining more than one trial
  - Who is likely to be non-compliant or drop out?
  - What are alternative (I.E., Non-approved) uses supported by the data?

Pharmaceutical Applications



- **High throughput screening**
  - Predict actions in assays
  - Predict results in animals or humans
- **Rational drug design**
  - Relating chemical structure with chemical properties
  - Inverse regression to predict chemical properties from desired structure
- **DNA snips**
- **Genomics**
  - Associate genes with diseases
  - Find relationships between genotype and drug response (e.g., dosage requirements, adverse effects)
  - Find individuals most susceptible to placebo effect

Fraud and Terrorist Detection

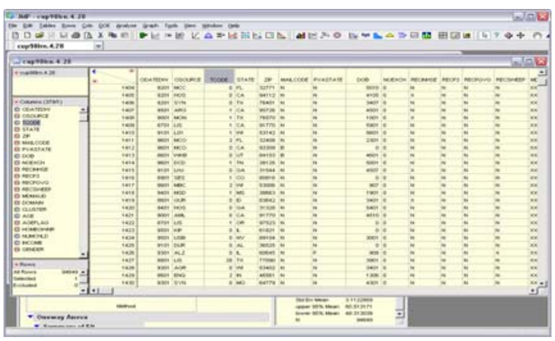
- **Identify false:**
  - Medical insurance claims
  - Accident insurance claims
- **Which stock trades are based on insider information?**
- **Whose cell phone numbers have been stolen?**
- **Which credit card transactions are from stolen cards?**
- **Which documents are "interesting"**
- **When are changes in networks signs of potential illegal activity?**

Lesson 1: Learn to Make Friends

- **PVA is a philanthropic organization,**
  - Sanctioned by the US Govt to represent the disabled veterans
- **They send out 4 million “free gifts” , every 6 weeks**
  - And hope for donations
- **Data were used for the KDD 1998 cup**
  - 200,000 donors
    - (100,000 training, 100,000 test)
  - 481 demographic variables
    - Past giving, income, age etc etc etc
  - Recent campaign (only for training set)
    - Did they give? (Target B)
    - How much did they give (Target D)

- **To optimize profit, who should receive the current solicitation?**
- **What is the most cost effective strategy?**

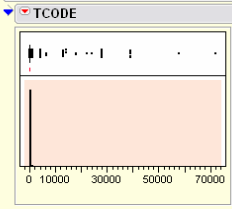
What’s “Hard”? --Example



T-Code

Distributions

TCODE



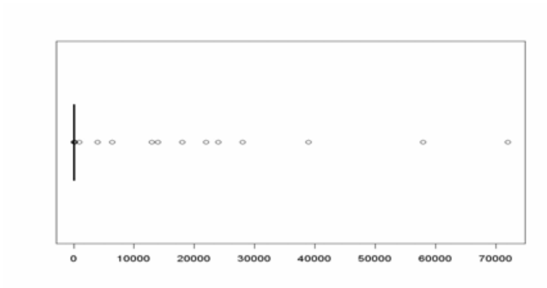
Quantiles

100.0%	maximum	72002
99.5%		1002
97.5%		28
90.0%		28
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Moments

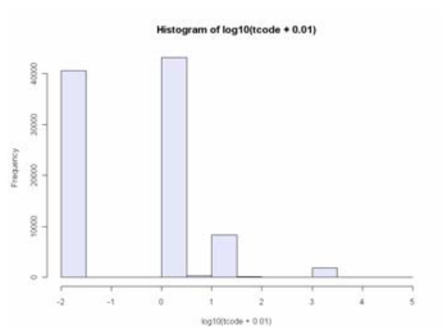
Mean	54.413105
Std Dev	957.50024
Std Err Mean	3.1122959
upper 95% Mean	60.513171
lower 95% Mean	48.313039
N	94649

More Tcode



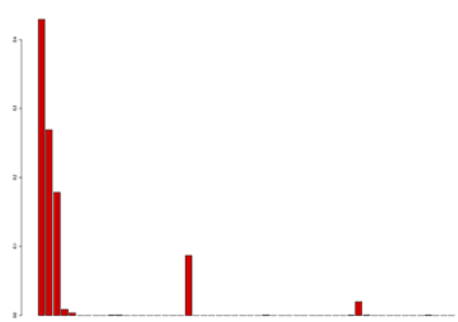
MAA Short Course

Transformation?



MAA Short Course

Categories?



MAA Short Course

What does it mean?

T-Code	Title						
0	_	16	DEAN	48	CORPORAL	109	LIC.
1	MR.	17	JUDGE	50	ELDER	111	SA.
1001	MESSRS.	17002	JUDGE & MRS.	56	MAYOR	114	DA.
1002	MR. & MRS.	18	MAJOR	59002	LIEUTENANT & MRS.	116	SR.
2	MRS.	18002	MAJOR & MRS.	62	LORD	117	SRA.
2002	MESDAMES	19	SENATOR	63	CARDINAL	118	SRTA.
3	MISS	20	GOVERNOR	64	FRIEND	120	YOUR MAJESTY
3003	MISSES	21002	SERGEANT & MRS.	65	FRIENDS	122	HIS HIGHNESS
4	DR.	22002	COLNEL & MRS.	68	ARCHDEACON	123	HER HIGHNESS
4002	DR. & MRS.	24	LIEUTENANT	69	CANON	124	COUNT
4004	DOCTORS	26	MONSIGNOR	70	BISHOP	125	LADY
5	MADAME	27	REVEREND	72002	REVEREND & MRS.	126	PRINCE
6	SERGEANT	28	MS.	73	PASTOR	127	PRINCESS
9	RABBI	28028	MSS.	75	ARCHBISHOP	128	CHIEF
10	PROFESSOR	29	BISHOP	85	SPECIALIST	129	BARON
10002	PROFESSOR & MRS.	31	AMBASSADOR	87	PRIVATE	130	SHEIK
10010	PROFESSORS	31002	AMBASSADOR & MRS	89	SEAMAN	131	PRINCE AND PRINCESS
11	ADMIRAL	33	CANTOR	90	AIRMAN	132	YOUR IMPERIAL MAJEST
11002	ADMIRAL & MRS.	36	BROTHER	91	JUSTICE	135	M. ET MME.
12	GENERAL	37	SIR	92	MR. JUSTICE	210	PROF.
12002	GENERAL & MRS.	38	COMMODORE	100	M.		
13	COLONEL	40	FATHER	103	MLLE.		
13002	COLONEL & MRS.	42	SISTER	104	CHANCELLOR		
14	CAPTAIN	43	PRESIDENT	106	REPRESENTATIVE		
14002	CAPTAIN & MRS.	44	MASTER	107	SECRETARY		
15	COMMANDER	46	MOTHER	108	LT. GOVERNOR		
15002	COMMANDER & MRS.	47	CHAPLAIN				

Relational Data Bases

- **Data are stored in tables**

```
Items
ItemID        ItemName      price
C56621        top hat       34.95
T35691        cane           4.99
RS5292        red shoes     22.95

Shoppers
Person ID     person name   ZIPCODE       item bought
135366         Lyle         19103          T35691
135366         Lyle         19103          C56621
259835         Dick         01267          RS5292
```


Metadata

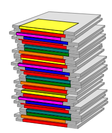
- **The data survey describes the data set contents and characteristics**
  - Table name
  - Description
  - Primary key/foreign key relationships
  - Collection information: how, where, conditions
  - Timeframe: daily, weekly, monthly
  - Cosynchronus: every Monday or Tuesday

Data Preparation

- **Build data mining database**
- **Explore data**
- **Prepare data for modeling**

60% to 95% of the time is spent preparing the data

Data Challenges



- **Data definitions**
  - Types of variables
- **Data consolidation**
  - Combine data from different sources
  - NASA mars lander
- **Data heterogeneity**
  - Homonyms
  - Synonyms
- **Data quality**

Missing Values

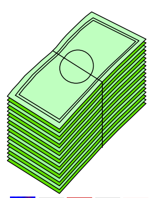
- **Random missing values**
  - Delete row?
    - Paralyzed Veterans
  - Substitute value
    - Imputation
    - Multiple Imputation
    - JMP 8 (!)
- **Systematic missing data**
  - Now what?

Missing Values -Systematic

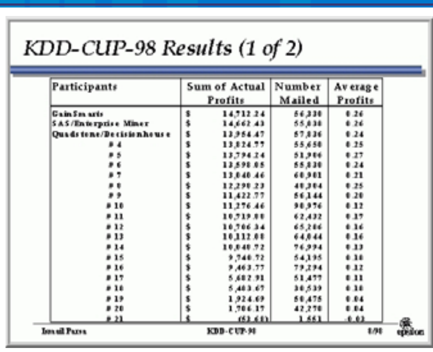
- Credit Card Bank finds that "Income" field is missing
- Wharton Ph.D. Student questionnaire on survey attitudes
- Bowdoin college applicants have mean SAT verbal score above 750
- Clinical Trial of Depression Medication – what does missing mean?

Results for PVA Data Set

- If entire list (100,000 donors) are mailed, net donation is $10,500
- Using data mining techniques, this was increased 41.37%



KDD CUP 98 Results

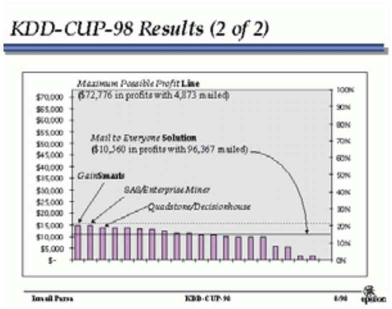


KDD-CUP-98 Results (1 of 2)

Participants	Sum of Actual Profits	Number Mailed	Average Profits
GainSmarts	$ 14,712.24	56,330	0.26
SAS/Enterprise Miner	$ 14,662.43	55,838	0.26
Quadstone/Decisionhouse	$ 13,954.47	57,836	0.24
# 4	$ 13,824.77	55,650	0.25
# 5	$ 13,794.24	51,906	0.27
# 6	$ 13,598.05	55,830	0.24
# 7	$ 13,040.46	60,901	0.21
# 8	$ 12,298.23	48,304	0.25
# 9	$ 11,422.77	56,144	0.20
# 10	$ 11,276.46	90,976	0.12
# 11	$ 10,719.88	62,432	0.17
# 12	$ 10,706.34	65,286	0.16
# 13	$ 10,112.08	64,044	0.16
# 14	$ 10,048.72	76,994	0.13
# 15	$ 9,740.72	54,195	0.18
# 16	$ 9,463.77	79,294	0.12
# 17	$ 5,682.91	51,477	0.11
# 18	$ 5,483.67	30,539	0.18
# 19	$ 1,924.69	50,475	0.04
# 20	$ 1,706.17	42,270	0.04
# 21	$ (53.61)	1,551	-0.03

Ismail Parsa KDD-CUP-98 8/98 epsilon

KDD CUP 98 Results 2



Data Mining vs. Statistics

Large amount of data:

30,000,000 rows, 1000 columns — 1,000 rows, 30 columns

Data Collection

Happenstance Data — Designed Surveys, Experiments

Sample?

Why bother? We have big, parallel computers — You bet! We even get error estimates.

Reasonable Price for Sofware

$1,000,000 a year — $599 with coupon from Amstat News

Presentation Medium

PowerPoint, what else? — Overhead foils are still the best

Nice Place for a Meeting

Aspen in January, Maui in February,… — Dallas in August, Orlando in August, Philadelphia in August, D.C. in January…

Data Mining Vs. Statistics

- Exploration - Flexible models
- Prediction often most important
- Computation matters
- Results are actionable
- Variable selection and overfitting are problems

- Tests of Hypotheses
  - Particular model and error structure
- Understanding, confidence intervals
- Computation not critical
- Results are interesting
- Variable selection and model selection are still problems

Knowledge Discovery Process

- Define business problem
- Build data mining database
- Explore data
- Prepare data for modeling
- Build model
- Evaluate model
- Deploy model and results

Note: This process model borrows from
CRISP-DM: CRoss Industry Standard Process for Data Mining

Data Mining Myths

- Find answers to unasked questions
- Continuously monitor your data base for interesting patterns
- Eliminate the need to understand your business
- Eliminate the need to collect good data
- Eliminate the need to have good data analysis skills

Successful Data Mining

- The keys to success:
  - Formulating the problem
  - Using the right data
  - Flexibility in modeling
  - Acting on results
- Success depends more on the way you mine the data rather than the specific tool

Data Mining and OLAP

- **On-line analytical processing (OLAP): users deductively analyze data to verify hypothesis**
  - Descriptive, not predictive
- **Data mining: software uses data to inductively find patterns – models!**
  - Predictive or descriptive
- **Associations?**
  - Most associated variables in the census
  - Most associated variables in a supermarket
  - Assocation Rules

Why Models?

- **Beer and Diapers**
  - "In the convenience stores we looked at, on Friday nights, purchases of beer and purcha of diapers are highly associated"
  - Conclusions?
  - Actions?

Models

- **Models are:**
  - Powerful summaries for understanding
  - Used for exploration and prediction
- **Of course, models are not reality**
- **George Box**
  - "All models are wrong, but some are useful"
  - "Statisticians, like artists, have the bad habit of falling in love with their models".

Twyman's Law and Corollaries

- **"If it looks interesting, it must be wrong"**
- **De Veaux's Corollary 1 to Twyman's Law**
  - "If it's perfect, it's wrong"
- **De Veaux's Corollary 2 to Twyman's Law**
  - "If it isn't wrong, you probably knew it already

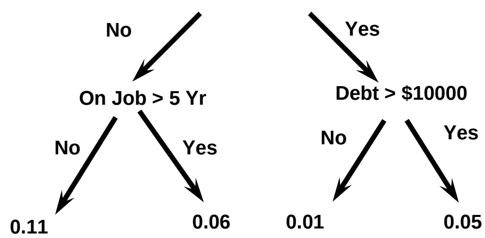
Lesson 2 – An Example of Twyman's Law

- Ingot cracking
  - 953 30,000 lb. Ingots
  - 8% cracking rate
  - $30,000 per recast
  - 90 potential explanatory variables
    - Water composition (reduced)
    - Metal composition
    - Process variables
    - Other environmental variables

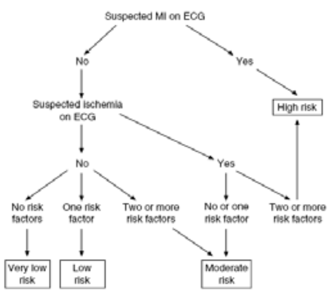
Data Processing

- **Five months to consolidate process data**
- **Three months to analyze and reduce dimension of water data**
- **Eight months after starting projects, statisticians received flat file:**
  - 960 ingots (rows)
  - 149 variables

Decision Trees – Mortgage Defaults



Cook County Hospital – "ER"



Confusion Matrix

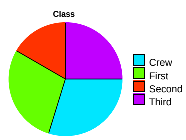
Doctors in ER	Actual Heart Attack	No Heart Attack
Predict Heart Attack	0.89	0.75
Predict No Heart Attack	0.11	0.25

Tree Algorithm (Goldman)	Actual Heart Attack	No Heart Attack
Predict Heart Attack	0.92	0.08
Predict No Heart Attack	0.04	0.96

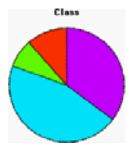
Two Way Tables -- Titanic

		Ticket Class				
		Crew	First	Second	Third	Total
	Lived	212	202	118	178	**710**
Survival	Died	673	123	167	528	**1491**
	Total	**885**	**325**	**285**	**706**	**2201**

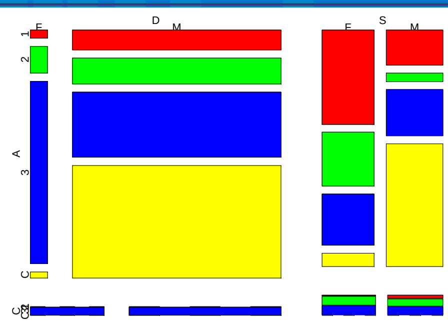
Survivors



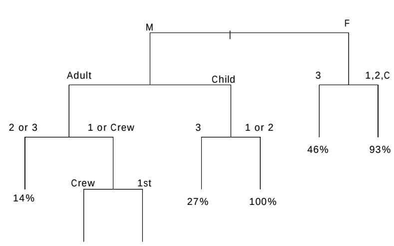
Non-Survivors



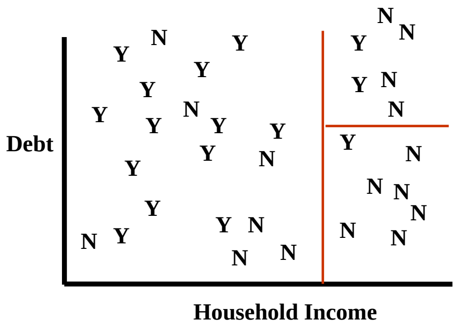
Mosaic Plot



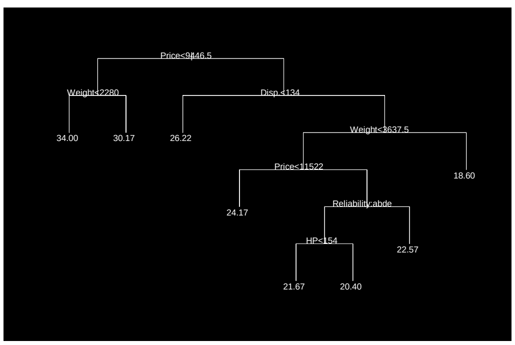
Tree Model



Geometry of Decision Trees

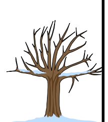


Regression Tree



Decision Trees -- Summary

- **Find split in predictor variable that best splits data into heterogeneous groups**
- **Build the tree inductively basing future splits on past choices (greedy algorithm)**
- **Classification trees (categorical response)**
- **Regression tree (continuous response)**
- **Size of tree often determined by cross-validation**



Tree Advantages

- **Model explains its reasoning -- builds rules**
- **Build model quickly**
- **Handles non-numeric data**
- **No problems with missing data**
  - Missing data as a new value
  - Surrogate splits
- **Works fine with many dimensions**

What's Wrong With Trees?

- **Output are step functions – big errors near boundaries**
- **Greedy algorithms for splitting – small changes change model**
- **Uses less data after every split**
- **Model has high order interactions -- all splits are dependent on previous splits**
- **Often non-interpretable**

Trees and Missing Values

- **Three advantages of trees**
  1. Can mix continuous and categorical predictors
  2. Selects subsets of predictors easily
  3. Can treat missing values as another category

First Tree

All Rows	
Count	912
Mean	8.1560673
Std Dev	18.003357

Alloy (6045,7348,8234,2345,3234)		Alloy (5434,5894,2439)	
Count	697	Count	215
Mean	4.6580583	Mean	19.496124
Std Dev	12.311747	Std Dev	26.79083

We know that – some alloys are hard to make. That's why we gave you the data in the first place.

Second Tree

All Rows	
Count	912
Mean	8.1560673
Std Dev	18.003357

CR<2.72		CR>=2.72	
Count	680	Count	232
Mean	4.9877451	Mean	17.442529
Std Dev	13.444396	Std Dev	25.115348

What do you think is *in* those alloys?

One More Time

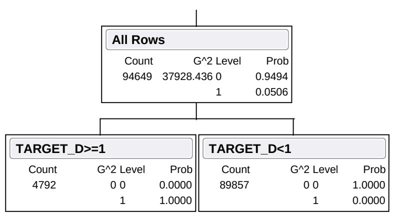
- **Looks like Manganese matters**
  - OH!
  - Did that solve it?
    - Experimental design
    - Enabled us to *focus* on important variables

I mean "Hmm.. That's funny."
-Issac Asimov

What did we learn?

- Data mining gave clues for generating hypotheses
- Followed up with DOE
- DOE led to substantial process improvement

Herb's Tree – Twyman's Law again

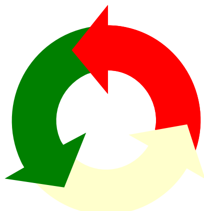


Types of Models

- Descriptions
- Classification (categorical or discrete values)
- Regression (continuous values)
  - Time series (continuous values)
- Clustering
- Association

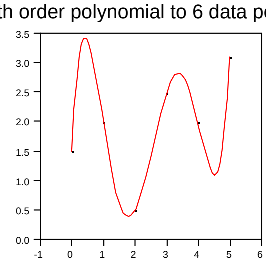
Model Building

- **Model building**
  - Train
  - Test
- **Evaluate**



Overfitting in Regression

Classical overfitting:
- Fit 6th order polynomial to 6 data points

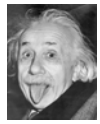


Overfitting

- **Fitting non-explanatory variables to data**
- **Overfitting is the result of**
  - Including too many predictor variables
  - Lack of regularizing the model
    - Neural net run too long
    - Decision tree too deep

Avoiding Overfitting

- **Avoiding overfitting is a balancing act – Occam's Razor**
  - Fit fewer variables rather than more
  - Have a reason for including a variable (other than it is in the database)
  - Regularize (don't overtrain)
  - Know your field.



All models should be as simple as possible but no simpler than necessary
Albert Einstein

Evaluate the Model

- **Accuracy**
  - Error rate
  - Proportion of explained variation
- **Significance**
  - Costs (symmetric?)
  - Statistical
  - Reasonableness
  - Sensitivity
  - Compute value of decisions
    - The "so what" test

Simple Validation

- ***Method :*** **split data into a training data set and a testing data set. A third data set for validation may also be used**
- ***Advantages:*** **easy to use and understand. Good estimate of prediction error for reasonably large data sets**
- ***Disadvantages*****: lose up to 20%-30% of data from model building**

Training vs. Test Data Sets

Train

Age	Income	Job Yrs	OK
41	29,000	8	Y
32	54,000	5	Y
26	29,000	2	N

Test

Age	Income	Job Yrs	OK	Model
39	29,000	4	Y	N
29	54,000	5	Y	Y

N-fold Cross Validation

- **Divide the data into N equal sized groups and build a model on the data with one group left out.**
- **Repeat for either systematic or random subgroups**
- **For very small data sets, N can be 1 (jackknife)**

1	2	3	4	5	6	7	8	9	10

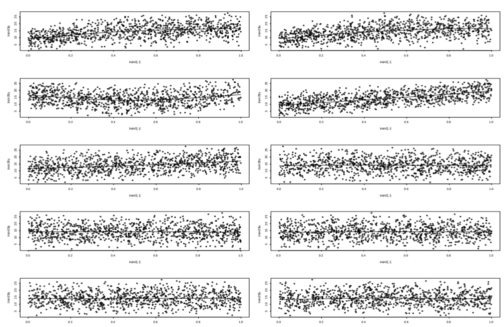
Internal vs. External Validation

- **Internal**
  - Many methods use internal cross-validation to choose size of model
    - Trees – depth
    - Regression – number of variables
    - Neural Network – size and type of architecture
- **External**
  - Used for model comparison
    - Compute R^2 on the test set
      - Just the correlation of predicted and actual
    - Compute confusion matrix on the test set

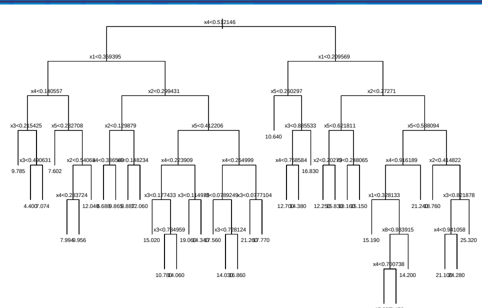
Regularization

- A model can be built to closely fit the training set but not the real data.
- Symptom: the errors in the training set are reduced, but increased in the test or validation sets.
- Regularization minimizes the residual sum of squares adjusted for model complexity.
- Accomplished by using a smaller decision tree or by pruning it. In neural nets, avoiding over-training.

“Toy” Problem

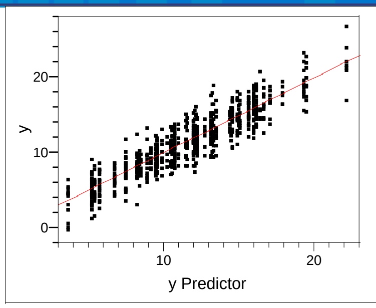


Tree Model



R –squared 82.3% Train 67.2% Test

Predictions for Example



R –squared 82.3% Train 67.2% Test

Linear Regression

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.900	0.482	-1.860	0.063
x1	4.658	0.292	15.950	<.0001
x2	4.685	0.294	15.920	<.0001
x3	-0.040	0.291	-0.140	0.892
x4	9.806	0.298	32.940	<.0001
x5	5.361	0.281	19.090	<.0001
x6	0.369	0.284	1.300	0.194
x7	0.001	0.291	0.000	0.998
x8	-0.110	0.295	-0.370	0.714
x9	0.467	0.301	1.550	0.122
x10	-0.200	0.289	-0.710	0.479

R-squared: 73.5% Train 69.4% Test

Stepwise Regression

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		-0.625	0.309	-2.019	0.0439
x1		4.619	0.289	15.998	<.0001
x2		4.665	0.292	15.984	<.0001
x4		9.824	0.296	33.176	<.0001
x5		5.366	0.28	19.145	<.0001

R-squared 73.4% on Train 69.8% Test

Stepwise 2ND Order Model

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.026	0.264	-7.68	<.0001
x1	4.311	0.184	23.47	<.0001
x2	4.808	0.185	26.04	<.0001
x3	-0.506	0.181	-2.79	0.0054
x4	10	0.186	53.79	<.0001
x5	5.212	0.176	29.67	<.0001
x8	-0.181	0.186	-0.97	0.3301
x9	0.427	0.188	2.28	0.0232
(x1-0.51811)*(x1-0.51811)	-0.932	0.711	-1.31	0.1905
(x2-0.48354)*(x1-0.51811)	8.972	0.634	14.14	<.0001
(x3-0.48517)*(x1-0.51811)	-1.367	0.65	-2.1	0.0358
(x3-0.48517)*(x2-0.48354)	-0.8	0.639	-1.25	0.2111
(x3-0.48517)*(x3-0.48517)	20.515	0.69	29.71	<.0001
(x4-0.49647)*(x1-0.51811)	1.014	0.651	1.56	0.1197
(x4-0.49647)*(x2-0.48354)	-1.159	0.65	-1.78	0.075
(x5-0.50509)*(x2-0.48354)	-0.794	0.62	-1.28	0.2008
(x5-0.50509)*(x3-0.48517)	1.105	0.619	1.78	0.0748
(x5-0.50509)*(x4-0.49647)	0.127	0.635	0.2	0.8414
(x8-0.52029)*(x5-0.50509)	1.065	0.63	1.69	0.0914

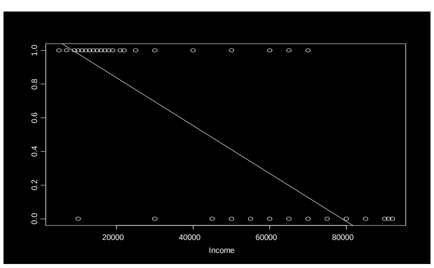
R-squared 89.9% Train 88.8% Test

Next Steps

- **Higher order terms?**
- **When to stop?**
- **Transformations?**
- **Too simple: underfitting – bias**
- **Too complex: inconsistent predictions, overfitting – high variance**
- **Selecting models is Occam's razor**
  - Keep goals of interpretation vs. prediction in mind

Logistic Regression

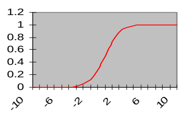
What happens if we use linear regression on 1-0 (yes/no) data?



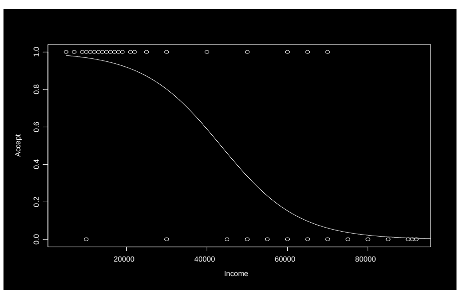
Logistic Regression II

- Points on the line can be interpreted as probability, but don't stay within [0,1]
- Use a sigmoidal function instead of linear function to fit the data

$$f(I)=\frac{1}{1+e^{-I}}$$



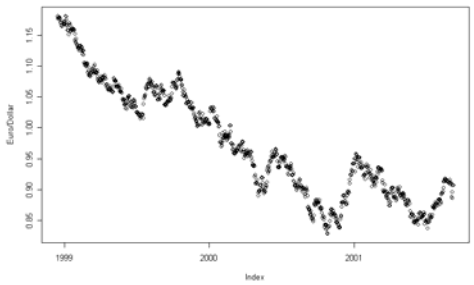
Logistic Regression III



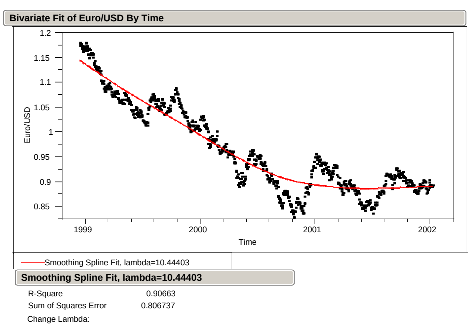
Regression - Summary

- Often works well
- Easy to use
- Theory gives prediction and confidence intervals
- Key is variable selection with interactions and transformations
- Use logistic regression for binary data

Smoothing – What's the Trend?

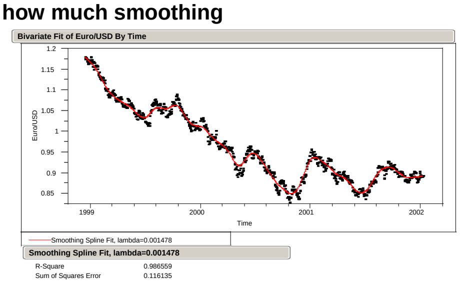


Scatterplot Smoother

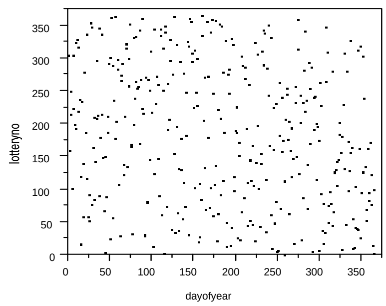


Less Smoothing

Usually these smoothers have choices on how much smoothing

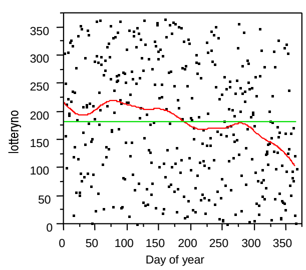


Draft Lottery 1970



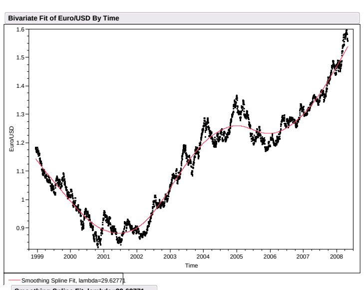
MAA Short Course January 3, 2009

Draft Data Smoothed



MAA Short Course January 3, 2009

Today



Smoothing Spline Fit, lambda=29.62771

R-Square	0.970806
Sum of Squares Error	3.250016

MAA Short Course January 3, 2009

More Dimensions

- **Why not smooth using 10 predictors?**
  - Curse of dimensionality
  - With 10 predictors, if we use 10% of each as a neighborhood, how many points do we need to get 100 points in cube?
  - Conversely, to get 10% of the points, what percentage do we need to take of each predictor?
  - Need new approach

Additive Model

- **Cant get**

$$\hat{y} = f\,(x_1, \ldots, x_p)$$

- **So, simplify to:**

$$\hat{y} = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

- **Each of the fi are easy to find**
  - Scatterplot smoothers

Create New Features

- **Instead of original x's use linear combinations**

$$z_i = \alpha + b_1 x_1 + \ldots + b_p x_p$$

  - Principal components
  - Factor analysis
  - Multidimensional scaling

How To Find Features

- **If you have a response variable, the question may change.**
- **What are interesting directions in the predictors?**
  - High variance directions in X - PCA
  - High covariance with Y -- PLS
  - High correlation with Y -- OLS
  - Directions whose smooth is correlated with *y* - PPR

Examples

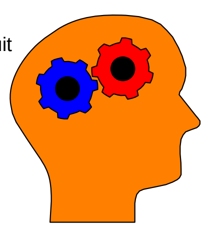
- **If the f's are sigmoidal (definition), this is called a neural network**

$$\hat{y} = \alpha + b_1 s_1(z_1) + b_2 s_2(z_2) + \ldots + b_p s_p(z_p)$$

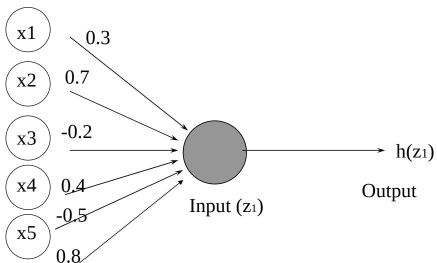
  - The z's are the hidden nodes
  - The s's are the activation functions
  - The b's are the weights

Neural Nets

- **Don't resemble the brain**
  - Are a statistical model
  - Closest relative is projection pursuit regression



A Single Neuron



$z_1 = 0.8 + .3x_1 + .7x_2 - .2x_3 + .4x_4 - .5x_5$

Single Node

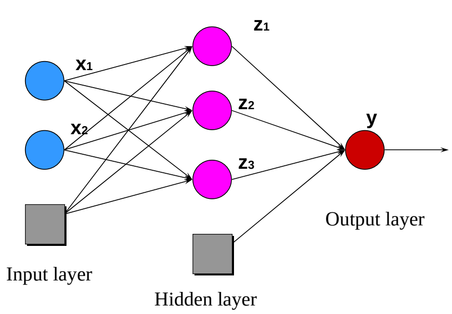
Input to outer layer from "hidden node":

$$I = z_l = \sum_j w_{1jk} x_j + \theta_l$$

Output:

$$\hat{y}_k = h(z_{kl})$$

Layered Architecture



Neural Networks

Create lots of features – hidden nodes

$$z_l = \sum_j w_{1jk} x_j + \theta_l$$

Use them in an additive model:

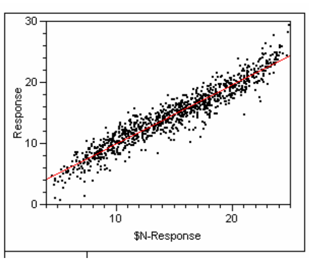
$$\hat{y}_k = w_{21}\ h(z_1) + w_{22}\ h(z_2) + \ldots + \theta_j$$

Put It Together

$$\hat{y}_k = \tilde{h}(\sum_l w_{2kl}\ \ h(\sum_j w_{1jk} x_j + \theta_l) + \theta_j)$$

The resulting model is just a flexible non-linear regression of the response on a set of predictor variables.

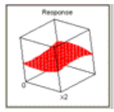
Predictions for Example

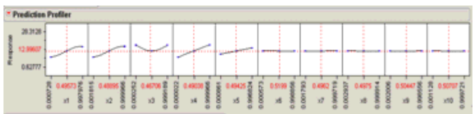


R^2 89.5% Train 87.7% Test

What Does This Get Us?

- **Enormous flexibility**
- **Ability to fit anything**
  - Including noise
- **Interpretation?**





Neural Net Pro

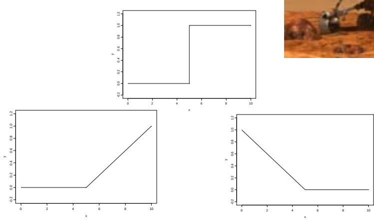
- **Advantages**
  - Handles continuous or discrete values
  - Complex interactions
  - In general, highly accurate for fitting due to flexibility of model
  - Can incorporate known relationships
    - So called grey box models
    - See De Veaux et al, *Environmetrics* 1999

Neural Net Con

- **Disadvantages**
  - Model is not descriptive (black box)
  - Difficult, complex architectures
  - Slow model building
  - Categorical data explosion
  - Sensitive to input variable selection

MARS – Multivariate Adaptive Regression Splines

- **What do they do?**
  - Replace each step function in a tree model by a pair of linear functions.



How Does It Work?

- **Replace each step function by a pair of linear basis functions.**
- **New basis functions may or may not be dependent on previous splits.**
- **Replace linear functions with cubics after backward deletions.**

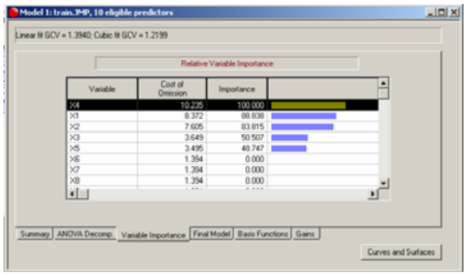
MARS Output

MARS modeling, version 3.5 (6/16/91)

forward stepwise knot placement:

basfn(s)		gcv	#indbsfns	#efprms	var	knot	parent
0		25.67	0.0	1.0			
1		17.36	1.0	7.0	4.	0.9308E-02	0.
3	2	12.26	3.0	14.0	1.	0.7059	0.
5	4	7.794	5.0	21.0	2.	0.6765	0.
7	6	6.698	7.0	28.0	3.	0.6465	1.
9	8	5.701	9.0	35.0	5.	0.3413	0.
11	10	5.324	11.0	42.0	1.	0.3754	4.
13	12	5.052	13.0	49.0	3.	0.3103	5.
15	14	5.869	15.0	56.0	4.	0.3269	2.
17	16	6.998	17.0	63.0	1.	0.5097	5.
19	18	8.761	19.0	70.0	3.	0.4290	0.
21	20	11.59	21.0	77.0	3.	0.8270	3.
23	22	20.83	23.0	84.0	3.	0.5001	2.
25	24	58.24	25.0	91.0	10.	0.2250	9.
26		461.7	26.0	97.0	10.	0.4740E-02	8.

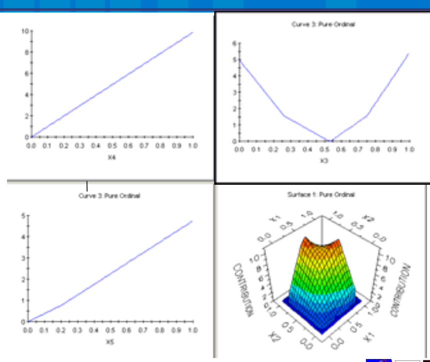
MARS Variable Importance



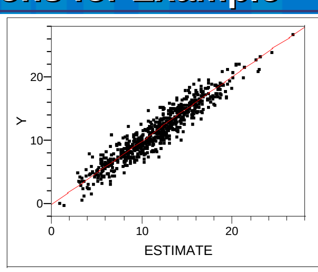
Variable	Cost of Omission	Importance
X4	10.235	100.000
X1	8.372	88.838
X2	7.605	83.815
X3	3.649	50.507
X5	3.495	48.747
X6	1.394	0.000
X7	1.394	0.000
X8	1.394	0.000

R-squared 95.0% Train 94.3% Test

MARS Function Output



Predictions for Example



R^2 = 95.0% Training Set 94.3% Test Set

Summary of MARS Features

- Produces smooth surface as a function of many predictor variables
- Automatically selects subset of variables
- Automatically selects complexity of model
- Tends to give low order interaction models preference
- Amount of smoothing and complexity may be tuned by user

K-Nearest Neighbors(KNN)

- **To predict *y* for an *x*:**
  - Find the *k* most similar *x*'s
  - Average their *y*'s
- **Find *k* by cross validation**
- **No training (estimation) required**
- **Works embarrassingly well**
  - Friedman, KDDM 1996

Collaborative Filtering

- Goal: predict what movies people will like
- Data: list of movies each person has watched

```
Lyle    André, Starwars
Ellen   André, Starwars, Destin
Fred    Starwars, Batman
Dean    Starwars, Batman, Rambo
Jason   Destin d'Amélie Poulin, Caché
```

Data Base

- Data can be represented as a sparse matrix

	Starwars	Rambo	Batman	My Dinner w/André	Destin D'Amilie	Caché
Lyle	y			y		
Ellen	y			y	y	
Fred	y	y				
Dean	y	y	y			
Jason	y			y		y
Karen	?	?	?	?	?	?

- Karen likes André. What else might she like?
- CDNow doubled e-mail responses

Clustering

- Turn the problem around
- Instead of predicting something about a variable, use the variables to group the observations
  - K-means
  - Hierarchical clustering

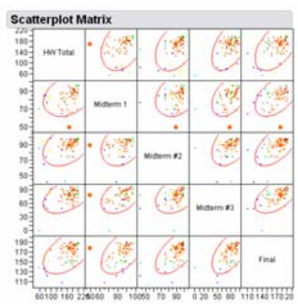
K-Means

- Rather than find the K nearest neighbors, find K clusters
- Problem is now to group observations into clusters rather than predict
- Not a predictive model, but a segmentation model

Example

- **Final Grades**
  - Homework
  - 3 Midterms
  - Final
- **Principal Components**
  - First is weighted average
  - Second is difference between 1 and 3rd midterms and 2nd

Scatterplot Matrix

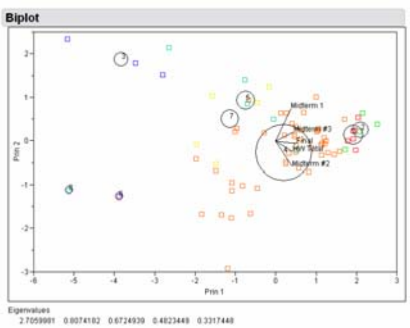


Cluster Means

Cluster Means

Cluster	HW Total	Midterm 1	Midterm #2	Midterm #3	Final
1	183.833333	91.3333333	97.8333333	93.3333333	188
2	195.75	94.75	98.25	89	188
3	81	83	61	59	130.333333
4	169.234043	82.7446809	91.2978723	77.8085106	172
5	172.2	86.2	75.8	81	151.2
6	139	65	84	50	110
7	84.4	85.2	90.8	72.4	164.4
8	56	71	87	0	139

Biplot



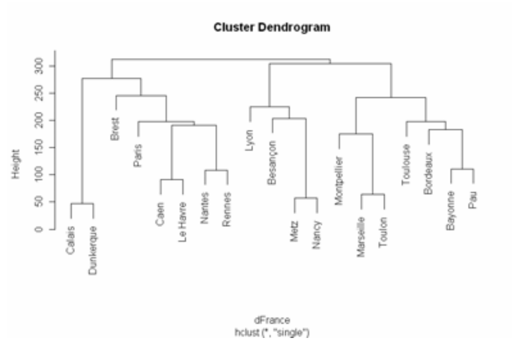
Hierarchical Clustering

- **Define distance between two observations**
- **Find closest observations and form a group**
  - Add on to this to form hierarchy

French Cities

	Bayonne	Besançon	Bordeaux	Brest	Caen	Calais	Dunkerqu	Le Havre	Lyon	Marseille	Metz	Montpelli	Nancy	Nantes	Paris	Pau	Rennes	Toulon	Toulouse
Bayonne	0	897	183	811	765	1056	1059	817	723	700	1088	537	1079	509	765	111	624	764	299
Besançon	897	0	695	962	636	606	608	598	225	536	263	528	204	747	405	951	714	601	758
Bordeaux	183	695	0	624	578	870	872	630	536	647	901	484	892	323	579	198	437	712	244
Brest	811	962	624	0	378	718	762	727	1014	1266	919	1103	891	298	596	817	246	1331	863
Caen	765	636	578	378	0	339	382	91	692	1004	573	928	556	295	236	771	191	1068	746
Calais	1056	606	870	718	339	0	47	278	750	1062	462	1053	481	598	290	1066	531	1126	970
Dunkerqu	1059	608	872	762	382	47	0	322	748	1060	440	1051	494	643	288	1064	575	1124	968
Le Havre	817	598	630	727	91	278	322	0	655	966	536	891	519	382	198	819	277	1031	819
Lyon	723	225	536	1014	692	750	748	655	0	314	456	305	404	661	458	728	766	378	535
Marseille	700	536	647	1266	1004	1062	1060	966	314	0	767	175	715	966	769	598	1014	64	405
Metz	1088	263	901	919	573	462	440	536	456	767	0	758	57	705	330	1097	672	832	988
Montpelli	537	528	484	1103	928	1053	1051	891	305	175	758	0	704	804	758	436	919	237	242
Nancy	1079	204	892	891	556	481	494	519	404	715	57	704	0	676	313	1130	643	780	937
Nantes	509	747	323	298	295	598	643	382	661	966	705	804	676	0	381	519	109	1033	565
Paris	765	405	579	596	236	290	288	198	458	769	330	758	313	381	0	773	348	834	678
Pau	111	951	198	817	771	1066	1064	819	728	598	1097	436	1130	519	773	0	632	663	198
Rennes	624	714	437	246	191	531	575	277	766	1014	672	919	643	109	348	632	0	1080	679
Toulon	764	601	712	1331	1068	1126	1124	1031	378	64	832	237	780	1033	834	663	1080	0	471
Toulouse	299	758	244	863	746	970	968	819	535	405	988	242	937	565	678	198	679	471	0

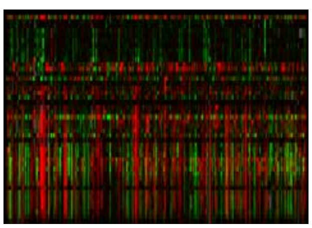
Dendogram



Series of experiments performed on the same set of genes

transcripts

experiments



44 experiments
x
407 genes

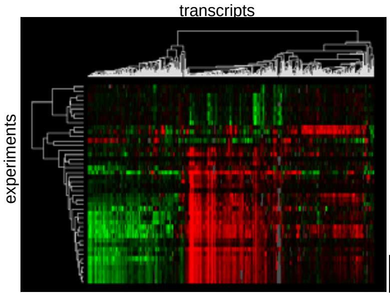
fold repression fold induction

*Eisen et al., 1998

Cluster the experiments and genes

transcripts

experiments



ste mutants

RHO O/X

PKC O/X

treatment
with
alpha-factor

fold repression fold induction

*data from Roberts et al., 2000

Grade Example

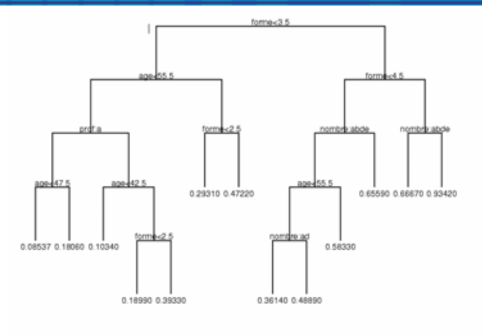
Lesson 3: Know When to Hold ‘em

- **Breast cancer data from mammograms**
  - Error rates by trained radiologists are near 25% for both false positives and false negatives
- **Newer equipment is prohibitively expensive for the developing world**
- **Early detection of breast cancer is crucial**
- **Cumulative type I error over a decade is near 100% leading to needless biopsies**

The Data

- **1618 mammograms showing clustered microcalcifications**
  - Biostatistics Dept Institut Curie
- **Variables**
  - Response: Malignant or not
  - Predictors: Age, Tissue Type (light/dense) Size (mm), Number of microcalc, Number of suspicious clusters, Shape of microcalc (1-5), Polyshape?(y/n), Shape of cluster (1,2,3), Retro (cluster near nipple?), Deep? (y/n)

Tree model



Combining Models

- **In 1950's forecasters found that combining forecasting models worked better on average than any single forecast model**
  - Reduces variance by averaging
  - Can reduce bias if collection is broader than single model

Bagging and Boosting

- **Bagging (Bootstrap Aggregation)**
  - Bootstrap a data set repeatedly
  - Take many versions of same model (e.g. tree)
    - Random Forest Variation
  - Form a committee of models
  - Take majority rule of predictions
- **Boosting**
  - Create repeated samples of weighted data
  - Weights based on misclassification
  - Combine by majority rule, or linear combination of predictions

Random Forests

- **Issues to decide**
  - How large a tree
  - How many trees
  - How many predictors to select at random for each tree
- **Breast cancer data**

MART

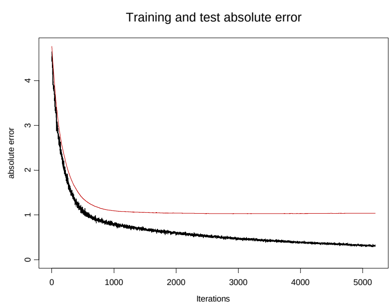
- **Boosting Version 1**
  - Use logistic regression.
  - Weight observations by misclassification
    - Upweight your mistakes
  - Repeat on reweighted data
  - Take majority vote
- **Boosting Version 2**
  - – use CART with 4-8 nodes
  - Use new tree on residuals
  - Repeat many, many times
  - Take predictions to be the sum of all these trees

Upshot of MART

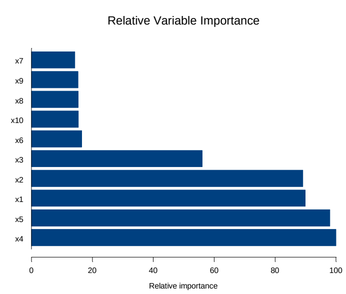
- **Robust – because of loss function and because we use trees**
- **Low interaction order because we use small trees (adjustable)**
- **Reuses all the data after each tree**

MART in action

More MART

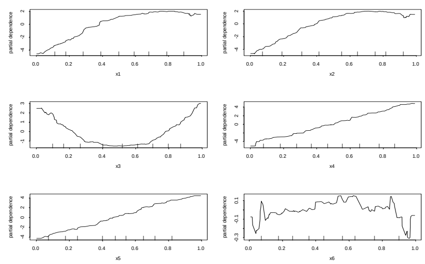


MART summary

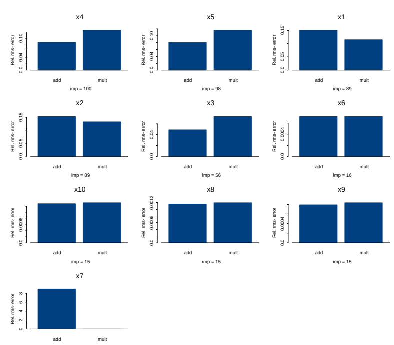


- R^2 --- 78.7 % on the test set.

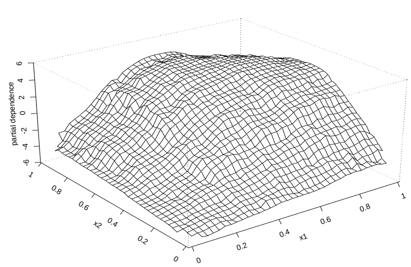
Single variable plots



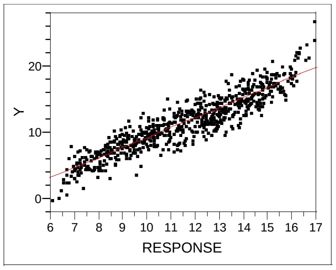
Interaction order?



Pairplots



MART Results



R squared 84.4% Train   78.7% Test

Results

- **Split data into train and test (62.5% - 37.5%)**
- **Repeat random splits 1000 times**
  - For each iteration, count false positives and false negatives on the 600 test set cases

	False Positives	False Negatives
Simple Tree	32.20%	33.70%
Neural Network	25.50%	31.70%
Boosted Trees	24.90%	32.50%
Bagged Trees	19.30%	28.80%
Radiologists	**22.40%**	**35.80%**

How Do We Really Start?

- **Life is not so kind**
  - Categorical variables
  - Missing data
  - 500 variables, not 10
- **481 variables – where to start?**

Where to Start

- **Three rules of data analysis**
  - Draw a picture
  - Draw a picture
  - Draw a picture
- **Ok, but how?**
  - There are 90 histogram/bar charts and 4005 scatterplots to look at (or at least 90 if you look only at y vs. X)

Exploratory Data Models

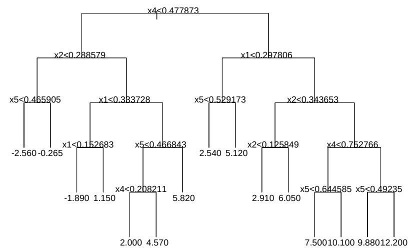
- **Use a tree to find a smaller subset of variables to investigate**
- **Explore this set graphically**
  - Start the modeling process over
- **Build model**
  - Compare model on small subset with full predictive model

More Realistic

- **250 predictors**
  - 200 Continuous
  - 50 Categorical
- **10,000 rows**
- **Why is this still easy?**
  - No missing values
  - Relatively high signal/noise

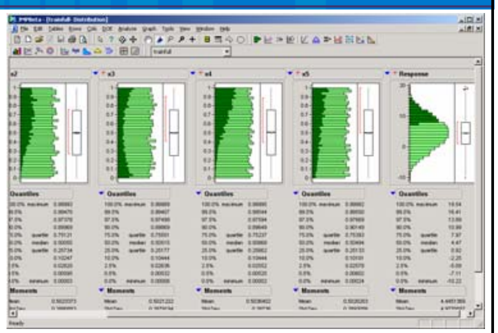
Start With a Simple Model

- Tree?



MAA Short Course

Brushing



MAA Short Course

Lesson 4: Know When to Fold 'em

- **Liability for churches**
  - Some Predictors
    - Net Premium Value
    - Property Value
    - Coastal (yes/no)
    - Inner100 (a.k.a., highly-urban) (yes/no)
    - High property value Neighborhood (yes/n
    - Indicator Class
      - 1 (Church/House of worship)
      - 2 (Sexual Misconduct – Church)
      - 3 (Add'l Sex. Misc. Covg Purchased)
      - 4 (Not-for-profit daycare centers)
      - 5 (Dwellings – One family (Lessor's risk))
      - 6 (Bldg or Premises – Office – Not for profit)
      - 7 (Corporal Punishment – each faculty member)
      - 8 (Vacant land- not for profit)
      - 9 (Private, not for profit, elementary, Kindergarten and Jr. High Schools)
      - 10 (Stores – no food or drink – not for profit)
      - 11 (Bldg or Premises – Maintained by insured (lessor's risk) – not for profit)
      - 12 (Sexual misconduct – diocese)

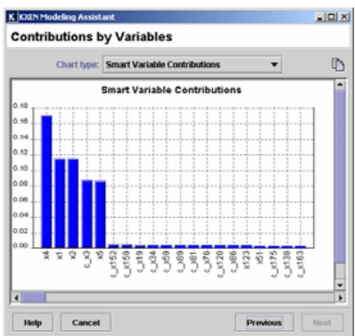
MAA Short Course

Fast Fail

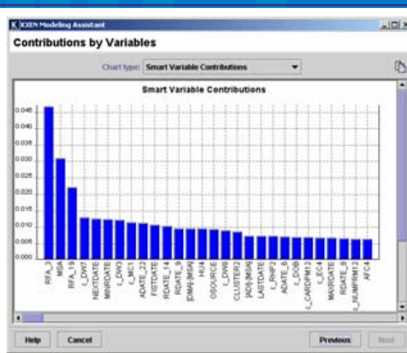
- Not every modeling effort is a success
  - A model search can save lots of queries
- Data took 8 months to get ready
- Analyst spent 2 months exploring it
- Tree models, stepwise regression (and a neural network running for several hours) found no out of sample predictive ability

Automatic Models

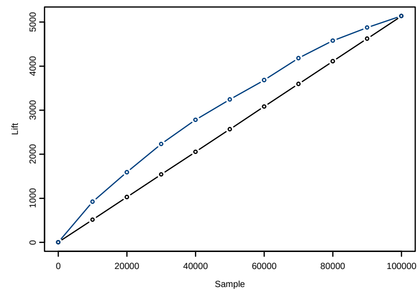
- KXEN



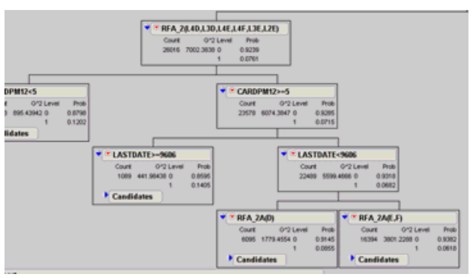
KXEN on PVA



Lift Curve



Exploratory Model



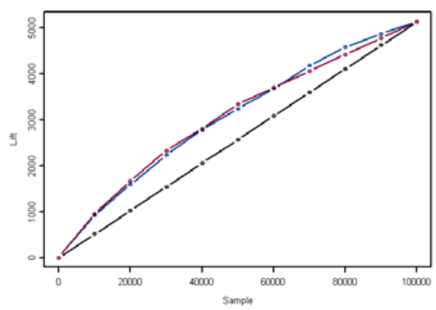
MAA Short Course January 3, 2009

Tree Model

- **Tree model on 40 key variables as indentified by KXEN**
  - Very similar performance to KXEN model
  - More coarse
  - Based only on
    - RFA_2
    - Lastdate
    - Nextdate
    - Lastgift
    - Cardprom

MAA Short Course January 3, 2009

Tree vs. KXEN



Lesson 5: Machines are Smart – You are Smarter

- **Why do statisticians like interpretability?**
- **Black boxes are not interpretable, but there may be important information**

Case Study – Warranty Data

- **A new backpack inkjet printer is showing higher than expected warranty claims**
  - What are the important variables?
  - What's going on?

- **A neural networks shows that Zip code is the most important predictor**

Zip Code?

Data Mining – DOE Synergy

- **Data Mining is exploratory**
- **Efforts can go on simultaneously**
- **Learning cycle oscillates naturally between the two**

What Did We Learn?

- **Toy problem**
  - Functional form of model
- **PVA data**
  - Useful predictor – increased sales 40%
- **Depression Study**
  - Identified critical intervention point at 2 weeks
- **Ingots**
  - Gave clues as to where to look
  - Experimental design followed
- **Churches**
  - When to quit
- **Printers**
  - When to experiment – what factors

Students in Stat 442

KDD-CUP-98 Results (1 of 2)

Participants		Sum of Actual Profits	Number Mailed	Average Profits
Student #1	$15,024	$ 14,712.24	56,330	0.26
Student #2	$14,695	$ 14,662.43	55,838	0.26
Student #3	$14,345	$ 13,954.47	57,836	0.24
		$ 13,824.77	55,650	0.25
# 5		$ 13,794.24	51,906	0.27
# 6		$ 13,598.05	55,830	0.24
# 7		$ 13,040.46	60,901	0.21
# 8		$ 12,298.23	48,304	0.25
# 9		$ 11,422.77	56,144	0.20
# 10		$ 11,276.46	90,976	0.12
# 11		$ 10,719.88	62,432	0.17
# 12		$ 10,706.34	65,286	0.16
# 13		$ 10,112.08	64,044	0.16
# 14		$ 10,048.72	76,994	0.13
# 15		$ 9,740.72	54,195	0.18
# 16		$ 9,463.77	79,294	0.12
# 17		$ 5,682.91	51,477	0.11
# 18		$ 5,483.67	30,539	0.18
# 19		$ 1,924.69	50,475	0.04
# 20		$ 1,706.17	42,270	0.04
# 21		$ (53.68)	1,551	-0.03

Ismail Parsa KDD-CUP-98 8/98 epsilon

Challenges for data mining

- Not algorithms
- Overfitting
- Finding an interpretable model that fits reasonably well

Recap – Success in Data Mining

- Problem formulation
- Data preparation
  - Data definitions
  - Data cleaning
  - Feature creation, transformations
- EDM – exploratory modeling
  - Reduce dimensions

Success in Data Mining II

- Don't forget Graphics
- Second phase modeling
- Testing, validation, implementation
- Constant re-evaluation of models

Which Method(s) to Use?

- No method is best
- Which methods work best when?
- Which method to use?
  - YES!

For More Information

- Two Crows
  - http//www.twocrows.com
- KDNuggets
  - http://www.kdnuggets.com
- deveaux@williams.edu

M. Berry and G. Linoff, **Data Mining Techniques,** Wiley, 1997

Dorian Pyle, **Data Preparation for Data Mining**, Morgan Kaufmann, 1999

Hand, D.J., Mannila, H. and Smyth, P., **Principles of Data Mining,** MIT Press 2001

Tan, P.N, Steinbach, and Kumar: **Introduction to Data Mining,** Addison-Wesley, 2006

Hastie, Tibshirani and Friedman, **Elements of Statistical Learning** 2nd edition, Springer